Knowledge Graph: Connecting Big Data Semantics

Ying Ding
Indiana University
Outline

• Vision
• Use Case: VIVO Ontology
• Use Case: Chem2Bio2RDF
• Challenges
VISION
Vision – Changes in Search

• *Strings vs. things*
Vision – Changes in Search

• **Relation** matters: connecting things/entities
Vision – Changes in Search

• Subgraph: Context is king
Vision – Changes in Search

• Future search:
  – string \(\rightarrow\) entity \(\rightarrow\) relation \(\rightarrow\) subgraph

• Filippo Menczer & Elinor Ostrom
Entities

- Entities are everywhere
- Entities on the Web: person, location, organization, book, music (vivoweb.org)
- Entities in medicine: gene, drug, disease, protein, side effect (chem2bio2rdf.org)
VIVO
VIVO: National networking of scientists

- VIVO: $12.5M funded by National Institute of Health to enable national networking of scientists
- 9/1/2009-8/31/2012, with one year extension
- 7 partners (Univ of Florida, Cornell Univ, Indiana University, Washington Univ, Scripps, Weill Cornell, Ponce Medical School)
- It utilizes Semantic Web technologies to model scientists and provides federated search to enhance the discovery of researchers and collaborators across the country
- Together with its sister project eagle-i ($13M), they will provide the semantic portals to network people and share resources.
Linked Open Data

DOAP SUMO DublineCore UMBEL SWRC Relationship DOLCE
FOAF SIOC Bibo AKT Biositemaps SUMO SWAN Popular Ontologies

Eagle-I Ontology U24 Ontologies VIVO Ontology

Biomedical Resource Ontology Information Artifact Ontology Neuroscience Information Framework Ontology for Biomedical Investigation Domain Ontologies
VIVO Ontology: Modeling Network of Scientists

• Network Structure:
  • People: foaf:Person, foaf:Organization,
  • Output: vivo:InformationResources
  • Relationship: vivo:role

• Academic Setting:
  – Research (bibo:Document, vivo:Grant, vivo:Project,
    vivo:Software, vivo:Dataset, vivo:ResearchLaboratory)
  – Teaching (vivo:TeacherRole, vivo:Course)
  – Service (vivo:Service, vivo:EditorRole, vivo:OrganizerRole, )
  – Expertise (skos:Concept)
Relationships have nuances

• The VIVO ontology supports representing rich information about relationships and how they change over time
  – description and duration of a person’s participation in a project or event
  – current and former employment, with titles and dates
  – author order in a publication

• Implemented as classes whose members we call context nodes
VIVO ontology localization

• Different localization required by different institutions
  – UF, Cornell, IU, WASHU, Scripps, MED-Cornell

• How to make localization:
  – Adding local namespace:
    • indiana: http://vivo.iu.edu/ontology/vivo-indiana/
    • core: http://vivoweb.org/ontology/core#
  – Local classes are the subclasses of the VIVO Core
    • foaf:Person → core:Non-academic→indiana:Professional Staff → indiana: AdministrativeServices
Modeling examples: Research

Modeling examples: Research

<http://vivo.iu.edu/individual/person25557>
rdf:type
<http://vivoweb.org/ontology/core#FacultyMember> .

<http://vivo.iu.edu/individual/person25557>
<http://vivoweb.org/ontology/core#authorInAuthorship>
<http://vivo.iu.edu/individual/n74> .

<http://vivo.iu.edu/individual/n74>
rdf:type
<http://vivoweb.org/ontology/core#Authorship> .

<http://vivo.iu.edu/individual/n74>
<http://vivoweb.org/ontology/core#linkedInformationResource>
<http://vivo.iu.edu/individual/n7109> .

<http://vivo.iu.edu/individual/n7109>
rdf:type
<http://purl.org/ontology/bibo/Article> .
Modeling examples: Research


<http://vivo.iu.edu/individual/n2881> <http://vivoweb.org/ontology/core#authorRank> 2 .

Applications

• Querying semantic data
  – SPARQL query builder

• Federated Search
  – VIVO Search
CHEM2BIO2RDF
Big Data in Life Sciences

• There is now an incredibly rich resource of public information relating compounds, targets, genes, pathways, and diseases. Just for starters there is in the public domain information on:
  – 69 million compounds and 449,392 bioassays (PubChem)
  – 59 million compound bioactivities (PubChem Bioassay)
  – 4,763 drugs (DrugBank)
  – 9 million protein sequences (SwissProt) and 58,000 3D structures (PDB)
  – 14 million human nucleotide sequences (EMBL)
  – 22 million life sciences publications - 800,000 new each year (PubMed)
  – Multitude of other sets (drugs, toxicogenomics, chemogenomics, metagenomics …)

• Even more important are the relationships between these entities. For example a chemical compound can be linked to a gene or a protein target in a multitude of ways:
  – Biological assay with percent inhibition, IC50, etc
  – Crystal structure of ligand/protein complex
  – Co-occurrence in a paper abstract
  – Computational experiment (docking, predictive model)
  – Statistical relationship
  – System association (e.g. involved in same pathways cellular processes)
How to take advantage of big data?

New biomedical insights

Knowledge discovery processes

Integrative Tools & Algorithms

Networks of data & relationships

Databases & Publications

Nuclear receptors:
- PPAR-gamma, PXR

SPARQL query builder
- Association Search & pathfinding
- ChemoHub: network predictive models
- Topic models & ranking
- WENDI & Chemogenomic Explorer
- Plotviz 3D visualization

Chem2Bio2RDF
- PubMedNet

Compounds, Drugs, Proteins,
- Genes, Pathways, Diseases,
- Side-Effects, Publications
need a data format!

need semantics!

http://chem2bio2rdf.org/drug/troglitazone

bindTo

http://chem2bio2rdf.org/target/PPARG
Chem2Bio2RDF

- NCI Human Tumor Cell Lines Data
- PubChem Compound Database
- PubChem Bioassay Database
- PubChem Descriptions of all PubChem bioassays
- Pub3D: A similarity-searchable database of minimized 3D structures for PubChem compounds
- Drugbank
- MRTD: An implementation of the Maximum Recommended Therapeutic Dose set
- Medline: IDs of papers indexed in Medline, with SMILES of chemical structures
- ChEMBL chemogenomics database
- KEGG Ligand pathway database
- Comparative Toxicogenomics Database
- PhenoPred Data

31m chemical structures
59m bioactivity data points
3m/19m publications
~5,000 drugs
Bio2RDF

Chem2Bio2RDF

LODD

Dereferenable URI

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemogenomics:ID</td>
<td><a href="http://www2cbio2rdf.org/resource/compound/3873%3E">http://www2cbio2rdf.org/resource/compound/3873&gt;</a></td>
</tr>
<tr>
<td>a:drugbank_interaction:DD_BENE</td>
<td><a href="http://www2cbio2rdf.org/resource/interaction/23120%3E">http://www2cbio2rdf.org/resource/interaction/23120&gt;</a></td>
</tr>
<tr>
<td>a:drugbank_interaction:DD_BENE</td>
<td><a href="http://www2cbio2rdf.org/resource/interaction/6320%3E">http://www2cbio2rdf.org/resource/interaction/6320&gt;</a></td>
</tr>
<tr>
<td>a:pubchem_browse:DD_BENE</td>
<td><a href="http://www2cbio2rdf.org/resource/pubchem_browse/43372%3E">http://www2cbio2rdf.org/resource/pubchem_browse/43372&gt;</a></td>
</tr>
<tr>
<td>chemogenomics:ID</td>
<td>PMID 3571</td>
</tr>
<tr>
<td>rdfs:label/</td>
<td>chemogenomics:3571_PMID</td>
</tr>
<tr>
<td>rdf:type/</td>
<td>vocab:chemogenomics</td>
</tr>
</tbody>
</table>

Browsing

RDF Triple store

SPARQL ENDPOINTS

PlotViz: Visualization

Cytoscape Plugin

Linked Path Generation and Ranking

RelFinder

Sigma

indice THE SEMANTIC INFORMATION MASHUP

Third party tools
Relating Pathways to Adverse Drug Reactions
RDF alone is not enough

• Need standardization

Troglitazone binds to PPARG

Romozins binds to PPARG

Romozins is another name of Troglitazone
Chem2Bio2OWL 1.0
<table>
<thead>
<tr>
<th>primary classes</th>
<th>description</th>
<th>sample instance data sources</th>
<th># of sample instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmallMolecule</td>
<td>a small bioactive molecule</td>
<td>PubChem, ChEBI</td>
<td>15509</td>
</tr>
<tr>
<td>Drug</td>
<td>a chemical used in the treatment, cure, prevention, or diagnosis of disease</td>
<td>DrugBank, PharmGKB, TTD</td>
<td>6544</td>
</tr>
<tr>
<td>Protein</td>
<td>a physical entity consisting of a sequence of amino acids</td>
<td>Uniprot, HGNC, GOA</td>
<td>12242</td>
</tr>
<tr>
<td>BioAssay</td>
<td>an experiment to measure the effects of some substance on target, cell, or a living organism</td>
<td>PubChem BioAssay, ChEMBL, BindingDB, DPSP</td>
<td>26861</td>
</tr>
<tr>
<td>Disease</td>
<td>any condition that causes pain, dysfunction, distress or social problems</td>
<td>OMIM, DO</td>
<td>8724</td>
</tr>
<tr>
<td>SideEffect</td>
<td>undesired effect from a medicine</td>
<td>SIDER</td>
<td>1385</td>
</tr>
<tr>
<td>Literature</td>
<td>a scientific article</td>
<td>Medline</td>
<td>28392</td>
</tr>
<tr>
<td>Pathway</td>
<td>a set or series of biological interactions</td>
<td>KEGG, Reactome</td>
<td>347</td>
</tr>
<tr>
<td>Drug-DrugInteraction</td>
<td>a drug affects the activity of another drug</td>
<td>DrugBank, DCDB</td>
<td>9690</td>
</tr>
<tr>
<td>Protein-Protein-Interaction</td>
<td>two or more proteins bind together</td>
<td>IPRD, DIP, BioGrid</td>
<td>54345</td>
</tr>
<tr>
<td>DrugInducedSideEffect</td>
<td>a drug interaction that results in side effect</td>
<td>SIDER</td>
<td>61102</td>
</tr>
<tr>
<td>DrugTreatmentChemicalProteinInteraction</td>
<td>the use of drug to treat disease genomic response to chemical compounds</td>
<td>Diseasesome, ChEMBL, BindingDB, DPSP, Ki, TTD, BindingMOAD, DrugBank, CTD, MATADOR, ArrayExpress, KEGG</td>
<td>47282</td>
</tr>
</tbody>
</table>
RDF Search
Target for Troglitazone

PREFIX c2b2r: http://chem2bio2rdf.org/chem2bio2rdf.owl#
PREFIX bp: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?uniprot_id
FROM <http://chem2bio2rdf.org/pubchem>
FROM <http://chem2bio2rdf.org/drugbank>
FROM <http://chem2bio2rdf.org/bindingdb>
FROM <http://chem2bio2rdf.org/uniprot>

WHERE {
    ?target bindingdb:MonomerId ?chemical.
    ?target bindingdb:IC50_value ?IC50 . FILTER (?IC50<10000) .
  )
  UNION
  )
} GROUP BY ?uniprot_id

FILTER (str(?drugName)="Troglitazone") 

Mashed Chem2Bio2RDF

Annotated Chem2Bio2OWL
SEMANTIC GRAPH MINING: PATH FINDING ALGORITHM

Dijkstra’s algorithm
Bio-LDA

- Latent Dirichlet Allocation (LDA)
  - The core of the group of powerful statistical modeling techniques for automated extraction of latent topics from large document collections

- Bio-LDA
  - Extended LDA model with Bio-terms as latent variable
  - Bio-terms: compound, gene, drug, disease, protein, side effect, pathways

- Calculate bio-term entropies over topics
- Use the Kullback-Leibler divergence as the non-symmetric distance measure for two bio-terms over topics
Example: Topic 10

| Topic Assignment | P(z|b) |
|------------------|-------|
| Venlafaxine      | 0.9676|
| HTR1A            | 0.9296|
| HTR2A            | 0.5871|
| Depressive Disorder | 0.9981 |
| Anxiety Disorder | 0.9445|
| Obsessive-Compulsive Disorder | 0.9782 |

Apply Bio-LDA on 336,899 PubMed article abstracts in 2009 and extract 50 topics
Diversity subgraph

Fig. Ranked association graphs between myocardial infarction and Troglitazone
Thiazolidinediones (TZDs) – revolutionary treatment for type II Diabetes

Troglitazone (Rezulin): withdrawn in 2000 (liver disease)

Rosiglitazone (Avandia): restricted in 2010 (cardiac disease)

Pioglitazone: ??? (does decrease blood sugar levels, was associated with bladder tumors and has been withdrawn in some countries.)

Rosiglitazone bound into PPAR-γ
**PPARG**: TZD target

**SAA2**: Involved in inflammatory response implicated in cardiovascular disease (Current Opinion in Lipidology 15,3,,269-278 2004)

**APOE**: Apolipoprotein E3 essential for lipoprotein catabolism. Implicated in cardiovascular disease.

**ADIPOQ**: Adiponectin involved in fatty acid metabolism. Implicated in metabolic syndrome, diabetes and cardiovascular disease

**CYP2C8**: Cytochrome P450 present in cardiovascular tissue and involved in metabolism of xenobiotics

**CDKN2A**: Tumor suppression gene

**SLC29A1**: Membrane transporter
Semantic Prediction
http://chem2bio2rdf.org/slap
Drug 1

- Substructure
- Side effect
- Chemical ontology
- Gene expression profile

Drug 2

Target 1

From Ligand perspective
From target perspective

Drug 1

bind

Target 1
- Sequence
- 3D structure
- GO
- Ligand

Target 2
Example: Troglitazone and PPARG

Troglitazone

Chemical ontology
bind

hypoglycemic drug

PPAR

bind

Eicosapentaenoic Acid

GO
pathway

Rosiglitazone

PPARG

bind

Pioglitazone

bind

Response to nutrient

GO
pathway

ACSL4

PPAR signaling pathway
Topology is important for association

Diagram:

- Cmpd 1 hasSubstructure Cmpd 2
- Cmpd 2 bind Protein 1

Diagram:

- Cmpd 1 hasSubstructure Cmpd 2
- Cmpd 2 bind Protein 1
Semantics is important for association.
SLAP Pipeline

(a) Raw Data Sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubChem</td>
<td></td>
</tr>
<tr>
<td>UniProt</td>
<td></td>
</tr>
<tr>
<td>SIDER</td>
<td></td>
</tr>
<tr>
<td>HPRD</td>
<td></td>
</tr>
<tr>
<td>BindingDB</td>
<td></td>
</tr>
<tr>
<td>CTD</td>
<td></td>
</tr>
<tr>
<td>DrugBank</td>
<td></td>
</tr>
<tr>
<td>UniProtKB-GOA</td>
<td></td>
</tr>
<tr>
<td>HGNC</td>
<td></td>
</tr>
<tr>
<td>OMIM</td>
<td></td>
</tr>
<tr>
<td>KEGG</td>
<td></td>
</tr>
<tr>
<td>ChEMBL</td>
<td></td>
</tr>
<tr>
<td>TTD</td>
<td></td>
</tr>
<tr>
<td>PDSP</td>
<td></td>
</tr>
</tbody>
</table>

(b) Ontological level schema

- Gene ontology
- Pathway
- Chemical ontology
- Disease
- Substructure

(c) Semantic Linked Data

(d) Paths (length <4) between Two Nodes
- Drug: Troglitazone
- Target: PPARG

(e) Significant Paths between Two Nodes
- Drug: Troglitazone
- Target: PPARG

(f) Statistical Models

1. Edge weight:
   \[ p(e(i \rightarrow j)) = \frac{1}{\sum_k R_{i,n} = R_{i,j}} \]
2. Path score:
   \[ p(P(t \rightarrow s)) = p(P) = \prod_{i=1}^{m-1} \epsilon_{i+1-i} \]
   \[ \log(p(P(t \rightarrow s))) = \sum_{i=1}^{m-1} \log(\epsilon_{i+1-i}) \]
3. Association score
   \[ \text{raw score}(s,t) = \sum_i \frac{\log(p(P(i)) - \theta(\log(P_i)))}{\sigma(\log(P_i))} \]
Cross-check with SEA

- SEA analysis (Nature 462, 175-181, 2009) predicts 184 new compound-target pairs, 30 of which were experimentally tested.
- 23 of these pairs were experimentally validated (<15uM) including 15 aminergic GPCR targets and 8 which crossed major receptor classification boundaries.
- 9 of the aminergic GPCR target pairings were correctly predicted by SLAP (p<0.05) – for the other 6 compounds were not present in our set.
- 1 of the 8 cross-boundary pairs was predicted.
Assessing drug similarity from biological function

- Took 157 drugs with 10 known therapeutic indications, and created SLAP profiles against 1,683 human targets
- Pearson correlation between profiles > 0.9 from SLAP was used to create associations between drugs
- Drugs with the same therapeutic indication unsurprisingly cluster together
- Some drugs with similar profile have different indications – potential for use in drug repurposing?
Challenges

• Generating entities: converting strings to things
  • Using **URI** to identify/integrate entities (RDF)
  • Using common schemas to represent **semantics** (ontologies)

• Managing relations:
  • Model **properties** of relations
  • **Search** and **rank** relations

• Handling context: tricky
  • Triples vs. Quads
  • Provenance: who says what, data-provenance, how (process)-provenance, workflow-provenance
Challenges

• Others:
  • Query efficiency,
  • Data security,
  • Data quality,
  • ...

Big Data + Big Challenge ➔ Unlimited Potential

Connect – Share – Discover
Thanks

dingying@indiana.edu
http://info.slis.indiana.edu/~dingying/index.html